HHS Public Access

Author manuscript

Curr Opin Psychol. Author manuscript; available in PMC 2019 December 01.

Published in final edited form as:

Curr Opin Psychol. 2018 December ; 24: 1–6. doi:10.1016/j.copsyc.2018.02.015.

The social neuroscience of mentalizing: challenges and recommendations

Dorit Kliemann and **Ralph Adolphs**

California Institute of Technology, Division of Humanities and Social Sciences, Pasadena, CA 91125

Abstract

Our ability to understand and think about the mental states of other people is referred to as "mentalizing" or "theory of mind". It features prominently in all social behavior, is essential for maintaining relationships, and shows pronounced individual differences. Here we review new approaches to study the underlying psychological mechanisms and discuss how they could best be investigated using modern tools from social neuroscience. We list key desiderata for the field, such as validity, specificity, and reproducibility, and link them to specific recommendations for the future. We also discuss new computational modeling approaches, and the application to psychopathology.

1. Introduction

Theory of mind (ToM), or mentalizing, refers to our ability to infer the hidden mental states of other people, such as their beliefs, intentions, and feelings. Mentalizing can be thought of as a specific kind of causal inference: inferring mental states that explain and predict people's behavior. The term "theory of mind" was initially coined as a term in connection with the question of whether chimpanzees might also have this ability [1, 2], but the field that provided most of the early empirical studies was developmental psychology. Studies in healthy children showed that explicit explanations and predictions related to the ability to understand other minds emerge around age four [3], although sensitivity to others' false beliefs may occur even earlier [4]. Atypical mentalizing has been suggested to lie at the core of social difficulties in autism, where its developmental emergence is notably delayed [5]. These findings fueled hypotheses about the cognitive and perceptual processes that might enable mentalizing in development [6], and about the social skills that it in turn enables, such as pretend play [7]. Across fields, challenges in research on this topic concern i) how to adequately measure behavioral nuances and quantify related neural substrates (measurement validity and precision), ii) how to approximate real life functioning in the laboratory (generalizability) and iii) how to inform psychopathology and personalized medicine

COI:

We have no conflict of interest.

(translational relevance). In this review, we provide recommendations for how best to make progress on these challenges, by enumerating a list of specific desiderata and how to achieve them.

2. List of desiderata

Validity

Mentalizing is nowadays tested with a variety of tasks ranging from the classic false belief attribution to social animations, rational action and strategic games [8]. While the choice of assessment depends highly on the research question and sample (animal, human, adult, infant), face, convergent, and discriminative validity are essential to build into all tasks that assess mentalizing. Since the term, “mentalizing,” is heterogeneously applied to begin with [8], tasks should be precise about what aspect of “mentalizing” or “ToM” they are supposed to measure (Implicit/explicit? Lexical or not? Representation or reasoning?). Ideally, there will be multiple tasks that can be used to converge on a putative aspect of this construct. Finally, tasks that do not require the inference of mental states, but that could be solved through associations between particular features and words (such as most tasks of emotion recognition, including the well-known “Mind in the Eyes” task [9] should not be taken as unambiguous markers of mentalizing without further evidence (see [10] for further discussion).

Studies in apes and young children are particularly challenging in their validity, because participants cannot be clearly instructed in the same way that adult humans can. Since posing the initial question [2] almost 40 years ago, a recent study in chimpanzees [**11] capitalized on a dependent measure also used with infants: preferential looking assessed with eyetracking. That study argued that great apes are indeed able to represent certain aspects of false beliefs. Yet there continues to be debate about whether or not nonhuman animals have ToM [12]. They certainly have the ability to predict aspects of another animal’s (or person’s) behavior. However, whereas adult humans can do so using relational representations that explicitly describe another’s mental state as decoupled from the state of the world, nonhuman animals may instead use a collection of other strategies. It has been argued that nonhuman primates use “awareness relations”, which flexibly represent what another individual is seeing or hearing [13]. While awareness relations provide considerable power in predicting behavior, they do not yet enable nonhuman animals to actively manipulate behavior. There is no good experimental evidence in animals yet of deception with the intent to cause a false belief, a behavior ubiquitous in humans.

Neuroscience studies may help resolve some of the unanswered questions, since they can provide independent evidence for shared or dissociable processes *(see Methodological Advances)*. For example, a recent experiment found that certain brain regions within the parietal and frontal lobes of the macaque brain, including parts of the dorsomedial prefrontal cortex (dMPFC; a region known to be involved in human mentalizing), were exclusively devoted to processing social interactions (but not object interactions) [**14]. A closely related study in human adults suggested a sensitivity to social interactions (but not the mere presence of two agents) within the posterior superior temporal sulcus (pSTS) [*15], a higher-order visual cortex region known to process biological motion. Given the accrual of

fine-grained neuroimaging data, comparisons across studies can now provide some differentiation of processes involved in mentalizing. It would be extremely valuable to obtain high-resolution neuroimaging data during social tasks in great apes, but so far this has been impossible for ethical and methodological reasons.

Specificity

There is general consensus for certain “social brain regions” involved in mentalizing [16–18]: bilateral temporoparietal junction (TPJ), precuneus (PC) and medial prefrontal cortex (MPFC). Several meta-analyses (e.g.,) have suggested that parts of this ‘core’ network are generally involved in mental state inferences, regardless of the task and stimulus formats. There remains a debate concerning the “specialization” of the mentalizing process, or its neural substrate. Perhaps consistent with early theoretical views that ToM arose as a form of social intelligence [2], some neuroimaging studies show considerable neural overlap between theory of mind, perspective taking [19], episodic memory and mental time travel [20]. A large part of the debate about the domain-specificity of mentalizing has centered on neuroimaging studies of the temporoparietal junction (TPJ) and dmPFC. The TPJ has been highlighted as selectively activated when we think about other people’s beliefs [21], and the most widely used functional magnetic resonance imaging (fMRI) localizer task for mentalizing uses an explicit lexical task requiring decisions about false beliefs-- and most selectively activates the TPJ [18, 22]. The specificity of the dMPFC in mentalizing remains debated [23, 24]: findings of amodal representations of emotional information in this region [25, 26] suggest a more general role in social cognition. Debates about whether these regions are specifically activated for ToM, or instead also subserve other functions [27, 28], will continue to be informed by better spatial resolution. Subregions, or neuronal subpopulations, can be anatomically indistinguishable, and their processes thus conflated, when using standard fMRI, even though they may well subserve different specialized functions [29–31]. It may also be advantageous to use functional localizer tasks that activate specific brain regions in individual subjects, rather than groups [32, 33] or to capitalize on new developments for aligning brains in group analyses on the basis of individual-level functional information [34, 35]. All of these developments should help provide better specificity of function through better anatomical precision.

Variability and individual differences

There are substantial individual differences in mentalizing abilities across adult individuals. Behavioral tasks in the laboratory often fail to incorporate this fact: few tasks show sufficient range in behavioral performance (but see, e.g.[36]). This poses a challenge how best to design neuroimaging experiments to capture the variability and explain individual differences in underlying neural responses. It further implies that group-level analyses need to be interpreted with caution, since they may reflect a heterogeneous mix of processes. Improvements are being made with the rapid development of more sensitive fMRI measures, as well as group analysis approaches that preserve individual functional differences [33–35], so that analyses can indeed be undertaken at the individual level.

The debate about the developmental trajectory of mentalizing abilities in children further underlines the importance to investigate individual differences. A comprehensive study in a

large sample of children and adults found functionally distinct mentalizing network responses already in 3-year-old children, i.e. before systematically passing explicit false belief task [37]. Older children also showed increased network specialization, overall arguing for a slow and continuous development of mentalizing.

Having sufficient statistical power even to detect individual differences is often a challenge. The typical sample sizes of neuroimaging studies ($n < 30$), for instance, rarely have sufficient power to capture reliable individual variability in functional MRI data. This could be partially addressed by future approaches that accrue large sets of publicly available fMRI data during metalizing tasks (see next section).

Methodological advances and reproducibility

Statistical reliability and reproducibility have received considerable attention recently in psychology and neuroscience in general [38]. Two recent examples of comprehensive replication studies investigated crucial psychological aspects of infant mentalizing [**39, 40]. General recommendations for increased reproducibility of fMRI research [**41] also apply to the neural mechanisms underlying mentalizing. Especially noteworthy in this context are new developments towards consistent preprocessing (e.g., fmriprep), data quality assessment (e.g., mriqc) and analyses (e.g., openneuro). Advances in imaging techniques, analyses, and atlases can further provide greater sensitivity and anatomical specificity. For instance, multi-echo fMRI together with independent components analysis has been reported to increase effect sizes in neuroimaging studies of mentalizing [42]. Atlases of subcortical structures crucial for social cognition, like the amygdala, can allow better distinction of activations within subnuclei [43, 44].

Multivariate analyses of fMRI data are now regularly used to identify specific representations of ToM, capitalizing on developments in machine learning algorithms. In these approaches, sensitivity to relative differences across spatial patterns of neural activity, rather than levels of mean activation, can improve the detection of mentalizing processes in general, as well as similarities and differences between individuals, groups [45, 46] and even species [47].

Finally, we suggest two additionally ingredients: 1) frequent pre-registration of all studies (which could be partial pre-registration), and 2) the accrual of data from mentalizing tasks into public shared databases.

Computational modeling

Recently, several computational models have been proposed for aspects of mentalizing, comprising predictive coding accounts [48, 49], learning models [50] or Bayesian-based inferences [**51, 52] (see Box for details on different approaches). These approaches connect with a large body of work using similar computational models in perception and decision-making. Formalizing psychological processes as computations may help to shed further light on the causal links between mentalizing, our own thoughts, and our social behavior. Assumptions and predictions about others' actions, plans, habits, desires, beliefs and ever-changing experience with the world have to be constantly updated and integrated.

A challenge common to most of these approaches is how to specifically connect the specific computations within the model to real neural computations.

BOX

Computational approaches to mentalizing

Game theory of mind models inferences of another agent's beliefs and goals to optimize our own behavior in mutual interactions. Here, goals represent different strategies, for instance operationalized within a social hunting game of two players [68]. These strategies are described as value-functions that take into consideration not only one's own behavior, but also that of another person. This in turn introduces different depths of recursion, which can be treated as *'levels of sophistication'* that each person has for thinking about what somebody else is thinking about [68]. In one study, individuals with autism showed difficulties inferring others' strategies [69], affecting their own optimal behavior in a game context. A different study, however, found no such strategic impairments during an economic game in autism [70].

Predictive coding accounts apply a theoretical framework from sensory perception to specific abstract mentalizing processes. Generally, predictive coding assumes that a response to a certain stimulus contains not only information about the stimulus itself, but additionally about the difference between the predicted state (or its value) and the actual state of the world. In other words, neural systems make forward predictions about expected information. Thus, predictions about others' intentions can be made, for instance, based on previous actions, group membership or personality aspects [48, 49]. Following this logic, activation in brain regions involved in ToM would be higher for unpredicted vs predicted mental states [71, 72].

Bayesian Theory of Mind (BToM) models aspects of ToM as Bayesian inference [51], by combining a generative model with a hypothesis space of possible mental states, and a prior over these hypotheses. The 'prior' represents the presumed probability of certain beliefs/desires, and the task then is to infer the posterior probability of unobservable mental states given our observations of behavior, an estimate that is continuously updated. Integrating and formalizing concepts of emotions within a Bayesian framework of ToM has recently used representational similarity analyses to link behavioral and neural data [*73].

Possible applications

BToM [51] attempts to "reverse engineer" others' mental state inferences (inspired by computational accounts of vision). It could be extended to more complex environments (than the ones tested so far) and has been applied to computations underlying mental state inferences in preverbal infants [74]. The predictive coding account holds promise for investigating psychopathology of social difficulties related to ToM: Autism has recently been hypothesized to reflect general as well as specific impairments in the ability to make predictions [75]. Both the predictive coding and BToM accounts also offer the challenge of how to account for the prior hypotheses in the first place, a relatively neglected domain

of study that will presumably bring us back to evolutionary considerations and comparative studies.

While there are differences in theoretical and mechanistic formalization between the various computational approaches, they together may offer a decisive step forward in trying to parse aspects of mentalizing into quantifiable (and predictable) components on the behavioral and neural level.

Applicability to psychopathology

Multiple psychiatric disorders show compromised mentalizing abilities (e.g. Borderline Personality Disorder [53], Schizophrenia [54] and autism [6]). General commonalities and specific differences in mentalizing across disorders are yet to be established. A first step will be to obtain a more comprehensive inventory of behavioral abnormalities across multiple tasks. For instance, adults with autism show difficulties in learning from social rewards (compared to monetary rewards) [55], show reduced social preferences when donating to charities [55], and are relatively insensitive to the presence of other people on tasks where others' opinions usually influence performance [56]. By contrast, the ability to perceive and extract emotional information from facial expressions [57, 58], animated social events and strategic levels of inference in competitive economic games [59] appear largely normal in the laboratory, even when manipulating the task relevance of social information [46]. Similarly, fMRI studies have found normal activation of the TPJ in autism during standard false-belief tasks [60], but atypical behavior when these are combined with inferences about moral responsibility [61].

These findings emphasize the need to carefully operationalize and reproducibly measure social abilities in general, and mentalizing in particular. Most useful for the field would be a standardized battery of lexical and nonlexical tasks that probe different aspects and uses of mentalizing. For instance, it would be valuable to use more naturalistic stimulus material [62, 63] in combination with tasks more closely approximating real life challenges (e.g., by adding temporal contingency [*59, 64] or an interaction component [65]).

Future Directions

Mentalizing tends to co-occur with other abilities into which it is sometimes decomposed. Which of these other abilities should be considered necessary components? Is there a causal relationship between mentalizing and these other abilities that we can elucidate (either phylogenetically, ontogenetically, or in ongoing cognition)? These are difficult questions since many perceptual and cognitive operations come into play during any actual realization of mentalizing (depending on the particular task used to assess it). It is difficult to disentangle which of these are mere enabling constraints and which are true components of the causal mechanisms behind mentalizing. Recent computational modeling approaches (see Box) together with advances in imaging methodology may help shed light on this question.

It would be important to obtain a broader inventory of which abilities correlate (and which do not) across behavioral as well as neural measures, just in healthy adults. For instance, while joint attention and mentalizing abilities appear to be highly correlated in infants, these

two abilities may not be correlated in adults [66] and are subserved by different neural substrates [67]. In the future, mentalizing will likely be parsed into more precise underlying psychological and neural computations that together comprise our theory of mind.

Acknowledgments

We thank David Amodio, Julian Jara-Ettinger, Hilary Richardson and Rebecca Saxe for comments on earlier versions of the manuscript.

Funded in part by the Caltech Conte Center for the Neurobiology of Social Decision-Making (NIMH).

References

1. Call J, Tomasello M. Does the chimpanzee have a theory of mind? 30 years later. Trends Cogn Sci. 2008; 12(5):187–92. [PubMed: 18424224]
2. Humphrey NK. The social function of intellect. In: Bateson PPG, Hinde RA, editorsGrowing points in ethology. Cambridge University Press; Cambridge, UK: 1976. 303–317.
3. Wellman HM, Cross D, Watson J. Meta-analysis of theory-of-mind development: The truth about false belief. Child Development. 2001; 72(3):655–684. [PubMed: 11405571]
4. Knudsen B, Liszkowski U. 18-Month-Olds Predict Specific Action Mistakes Through Attribution of False Belief, Not Ignorance, and Intervene Accordingly. Infancy. 2012; 17(6):672–691.
5. Baron-Cohen S, Leslie AM, Frith U. Does the autistic child have a “theory of mind”? Cognition. 1985; 21(1):37–46. [PubMed: 2934210]
6. Baron-Cohen S. Mindblindness: an Essay on Autism and Theory of Mind. Cambridge, MA: MIT Press; 1995.
7. Leslie AM. Pretense and Representation - the Origins of Theory of Mind. Psychological Review. 1987; 94(4):412–426.
8. Schaafsma SM, et al. Deconstructing and reconstructing theory of mind. Trends in Cognitive Sciences. 2015; 19(2):65–72. [PubMed: 25496670]
9. Baron-Cohen S, et al. The “Reading the Mind in the Eyes” Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. J Child Psychol Psychiatry. 2001; 42(2):241–51. [PubMed: 11280420]
10. Heyes C. Submentalizing: I Am Not Really Reading Your Mind. Perspect Psychol Sci. 2014; 9(2): 131–43. [PubMed: 26173251]
**11. Krupenye C, et al. Great apes anticipate that other individuals will act according to false beliefs. Science. 2016; 354(6308):110–114. Studying an impressive sample of great apes, this study found that apes could anticipate where human agents would look for a hidden item (although the apes knew the item was in a different location). The study was a novel application of using eyetracking in apes, and is one of the strongest to infer that apes have ToM (although this remains still debated). [PubMed: 27846501]
12. Spence CE, Osman M, McElligott AG. Theory of Animal Mind: Human Nature or Experimental Artefact? Trends in Cognitive Sciences. 2017; 21(5):333–343. [PubMed: 28347613]
13. Martin A, Santos LR. What Cognitive Representations Support Primate Theory of Mind? Trends in Cognitive Sciences. 2016; 20(5):375–382. [PubMed: 27052723]
**14. Sliwa J, Freiwald WA. A dedicated network for social interaction processing in the primate brain. Science. 2017; 356(6339):745–749. Neuroimaging responses to videos showing interactions between monkeys (social) and/or objects (nonsocial) revealed a distributed network of frontal and parietal brain regions disproportionately activated by social interactions in macaques. [PubMed: 28522533]
*15. Isik L, et al. Perceiving social interactions in the posterior superior temporal sulcus. Proc Natl Acad Sci U S A. 2017; 114(43):E9145–E9152. Within humans, a region within the posterior portion of STS was more activated by social interactions than by the mere presence of just two agents (without interaction). This region also contained information about the valence of the interaction (helping (positive) vs hindering (negative)). [PubMed: 29073111]

16. Aichhorn M, et al. Temporo-parietal Junction Activity in Theory-of-Mind Tasks: Falseness, Beliefs, or Attention. Journal of Cognitive Neuroscience. 2009; 21(6):1179–1192. [PubMed: 18702587]
17. Frith CD, Frith U. The Physiological Basis of Theory of Mind. In: Baron-Cohen S, Tager-Flusberg H, Cohen D, editorsUnderstanding Other Minds: Perspective From Developmental Social Neuroscience. Oxford University Press; Oxford: 2000. 335–356.
18. Saxe R, Kanwisher N. People thinking about thinking people - The role of the temporo-parietal junction in "theory of mind". Neuroimage. 2003; 19(4):1835–1842. [PubMed: 12948738]
19. Corbetta M, Shulman GL. Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci. 2002; 3(3):201–15. [PubMed: 11994752]
20. Buckner RL, Carroll DC. Self-projection and the brain. Trends Cogn Sci. 2007; 11(2):49–57. [PubMed: 17188554]
21. Saxe R, Powell LJ. It's the thought that counts: specific brain regions for one component of theory of mind. Psychol Sci. 2006; 17(8):692–9. [PubMed: 16913952]
22. Saxe R. Why and how to study Theory of Mind with fMRI. Brain Res. 2006; 1079(1):57–65. [PubMed: 16480695]
23. Amodio DM, Frith CD. Meeting of minds: the medial frontal cortex and social cognition. Nature Reviews Neuroscience. 2006; 7(4):268–277. [PubMed: 16552413]
24. Schurz M, et al. Clarifying the role of theory of mind areas during visual perspective taking: Issues of spontaneity and domain-specificity. Neuroimage. 2015; 117:386–96. [PubMed: 25907759]
25. Peelen MV, Atkinson AP, Vuilleumier P. Supramodal Representations of Perceived Emotions in the Human Brain. Journal of Neuroscience. 2010; 30(30):10127–10134. [PubMed: 20668196]
26. Skerry AE, Saxe R. A Common Neural Code for Perceived and Inferred Emotion. Journal of Neuroscience. 2014; 34(48):15997–16008. [PubMed: 25429141]
27. Carter RM, Huettel SA. A nexus model of the temporal-parietal junction. Trends Cogn Sci. 2013; 17(7):328–36. [PubMed: 23790322]
28. Mitchell JP. Activity in right temporo-parietal junction is not selective for theory-of-mind. Cereb Cortex. 2008; 18(2):262–71. [PubMed: 17551089]
29. Scholz J, et al. Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. PLoS One. 2009; 4(3):e4869. [PubMed: 19290043]
30. Mars RB, et al. Connectivity-Based Subdivisions of the Human Right "Temporoparietal Junction Area": Evidence for Different Areas Participating in Different Cortical Networks. Cerebral Cortex. 2012; 22(8):1894–1903. [PubMed: 21955921]
31. Rushworth MFS, Mars RB, Sallee J. Are there specialized circuits for social cognition and are they unique to humans? Current Opinion in Neurobiology. 2013; 23(3):436–442. [PubMed: 23290767]
32. Spunt RP, Adolphs R. Validating the Why/How contrast for functional MRI studies of Theory of Mind. Neuroimage. 2014; 99:301–311. [PubMed: 24844746]
33. Fedorenko E, et al. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J Neurophysiol. 2010; 104(2):1177–94. [PubMed: 20410363]
34. Haxby JV, et al. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron. 2011; 72(2):404–16. [PubMed: 22017997]
35. Guntupalli JS, et al. A Model of Representational Spaces in Human Cortex. Cereb Cortex. 2016; 26(6):2919–2934. [PubMed: 26980615]
36. Dziobek I, et al. Introducing MASC: A movie for the assessment of social cognition. Journal of Autism and Developmental Disorders. 2006; 36(5):623–636. [PubMed: 16755332]
37. Richardson H, et al. Development of the social brain from age three to twelve years. Nature communications. Nature communications. accected.
38. Aarts AA, et al. Estimating the reproducibility of psychological science. Science. 2015; 349(6251)
**39. Powell L, et al. Replications of implicit theory of mind tasks with varying representational demands. Cognitive Development. 2017 The authors tested (exact or conceptual) replications of experiments that previously reported mental state inferences in toddlers and infants and carefully discuss contributing factors and implications. Two experiments replicated all or some previously reported effects of implicit theory of mind in 2-year-olds and 18-month-olds. A third study did

not replicate an effect known as *violation of expectation,* suggesting limited infant false belief understanding in 18-month-olds.

40. Fizke E, et al. Are there signature limits n early theory of mind? Journal of Experimental Child Psychology. 2017; 162:209–224. [PubMed: 28623778]

**41. Nichols TE, et al. Best practices in data analysis and sharing in neuroimaging using MRI. Nature Neuroscience. 2017; 20(3):299–303. Based on work of the Committee on Best Practice in Data Analysis and Sharing (COBIDAS) of the Organization for Human Brain Mapping (OHBM) the authors put together benchmark guidelines to improve the standards of practice and reporting of MRI data. [PubMed: 28230846]

42. Lombardo MV, et al. Improving effect size estimation and statistical power with multi-echo fMRI and its impact on understanding the neural systems supporting mentalizing. Neuroimage. 2016; 142:55–66. [PubMed: 27417345]

43. Saygin ZM, et al. High-resolution magnetic resonance imaging reveals nuclei of the human amygdala: manual segmentation to automatic atlas. Neuroimage. 2017; 155:370–382. [PubMed: 28479476]

44. Tyszka JM, Pauli WM. In vivo delineation of subdivisions of the human amygdaloid complex in a high-resolution group template. Human Brain Mapping. 2016; 37(11):3979–3998. [PubMed: 27354150]

45. Coutanche MN, Thompson-Schill SL, Schultz RT. Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. Neuroimage. 2011; 57(1):113–123. [PubMed: 21513803]

46. Kliemann D, et al. Cortical responses to dynamic emotional facial expressions generalize across stimuli, and are sensitive to task-relevance, in adults with and without Autism. Cortex. (accepted).

47. Kriegeskorte N. Characterizing categorical and continuous visual-object codes in man, monkey, and computational models with representational similarity analysis. Perception. 2009; 38:62–62.

48. Kilner JM, Friston KJ, Frith CD. Predictive coding: an account of the mirror neuron system. Cogn Process. 2007; 8(3):159–66. [PubMed: 17429704]

49. Koster-Hale J, Saxe R. Theory of Mind: A Neural Prediction Problem. Neuron. 2013; 79(5):836–848. [PubMed: 24012000]

50. Keysers C, Gazzola V. Hebbian learning and predictive mirror neurons for actions, sensations and emotions. Philosophical Transactions of the Royal Society B-Biological Sciences. 2014; 369(1644)

51. Baker CL, et al. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nature Human Behavior. 2017:1. The authors propose a model formalizing core mentalizing computations as Bayesian inference about unobservable mental states and perceptions. The model was tested across two experiments and accurately captured mentalizing judgments about agents' goals and actions.

52. Baker CL, Saxe R, Tenenbaum JB. Action understanding as inverse planning. Cognition. 2009; 113(3):329–49. [PubMed: 19729154]

53. Dziobek I, et al. Neuronal correlates of altered empathy and social cognition in borderline personality disorder. Neuroimage. 2011; 57(2):539–548. [PubMed: 21586330]

54. Montag C, et al. Different aspects of theory of mind in paranoid schizophrenia: Evidence from a video-based assessment. Psychiatry Research. 2011; 186(2–3):203–209. [PubMed: 20947175]

55. Lin A, Rangel A, Adolphs R. Impaired learning of social compared to monetary rewards in autism. Front Neurosci. 2012; 6:143. [PubMed: 23060743]

56. Izuma K, et al. Insensitivity to social reputation in autism. Proc Natl Acad Sci U S A. 2011; 108(42):17302–7. [PubMed: 21987799]

57. Harms MB, Martin A, Wallace GL. Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. Neuropsychol Rev. 2010; 20(3):290–322. [PubMed: 20809200]

58. Kennedy DP, Adolphs R. Perception of emotions from facial expressions in high-functioning adults with autism. Neuropsychologia. 2012; 50(14):3313–9. [PubMed: 23022433]

*59. Pantelis PC, et al. A specific hypoactivation of right temporo-parietal junction/posterior superior temporal sulcus in response to socially awkward situations in autism. Social Cognitive and Affective Neuroscience. 2015; 10(10):1348–1356. Using temporally contingent movie stimuli

this functional neuroimaging study found reliable activation in the mentalizing network in both healthy subjects and those with autism. [PubMed: 25698698]

60. Dufour N, et al. Similar brain activation during false belief tasks in a large sample of adults with and without autism. PLoS One. 2013; 8(9):e75468. [PubMed: 24073267]
61. Moran JM, et al. Impaired theory of mind for moral judgment in high-functioning autism. Proc Natl Acad Sci U S A. 2011; 108(7):2688–92. [PubMed: 21282628]
62. Zaki J, Ochsner K. The need for a cognitive neuroscience of naturalistic social cognition. Ann N Y Acad Sci. 2009; 1167:16–30. [PubMed: 19580548]
63. Rosenblau G, et al. Approximating Implicit and Explicit Mentalizing with Two Naturalistic Video-Based Tasks in Typical Development and Autism Spectrum Disorder. Journal of Autism and Developmental Disorders. 2015; 45(4):953–965. [PubMed: 25267068]
64. Byrge L, et al. Idiosyncratic Brain Activation Patterns Are Associated with Poor Social Comprehension in Autism. Journal of Neuroscience. 2015; 35(14):5837–5850. [PubMed: 25855192]
65. Redcay E, et al. Live face-to-face interaction during fMRI: A new tool for social cognitive neuroscience. Neuroimage. 2010; 50(4):1639–1647. [PubMed: 20096792]
66. Shaw JA, et al. The relationship between joint attention and theory of mind in neurotypical adults. Consciousness and Cognition. 2017; 51:268–278. [PubMed: 28433857]
67. Deen B, et al. Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. Cerebral Cortex. 2015; 25(11):4596–4609. [PubMed: 26048954]
68. Yoshida W, Dolan RJ, Friston KJ. Game Theory of Mind. Plos Computational Biology. 2008; 4(12)
69. Yoshida W, et al. Cooperation and Heterogeneity of the Autistic Mind. Journal of Neuroscience. 2010; 30(26):8815–8818. [PubMed: 20592203]
70. Pantelis PC, Kennedy DP. Autism does not limit strategic thinking in the “beauty contest” game. Cognition. 2017; 160:91–97. [PubMed: 28081516]
71. Buckholtz JW, et al. The Neural Correlates of Third-Party Punishment. Neuron. 2008; 60(5):930–940. [PubMed: 19081385]
72. Koster-Hale J, et al. Decoding moral judgments from neural representations of intentions. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110(14): 5648–5653. [PubMed: 23479657]
*73. Saxe R, Houlihan SD. Formalizing emotion concepts within a Bayesian model of theory of mind. Curr Opin Psychol. 2017; 17:15–21. This theoretical review proposes formalizing of fine-grained emotion concepts in a Bayesian hierarchical generative model based on intuitive theories of other minds. [PubMed: 28950962]
74. Hamlin JK, et al. The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. Developmental Science. 2013; 16(2):209–226. [PubMed: 23432831]
75. Sinha P, et al. Autism as a disorder of prediction. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(42):15220–15225. [PubMed: 25288765]

HIGHLIGHTS

- Human social behavior requires understanding and thinking about other minds (mentalizing)
- Despite a large body of research across psychology and neuroscience, research on mentalizing faces challenges in validity, specificity and reproducibility
- We lay out desiderata and make recommendations for the field
- We review recent computational modeling approaches to mentalizing

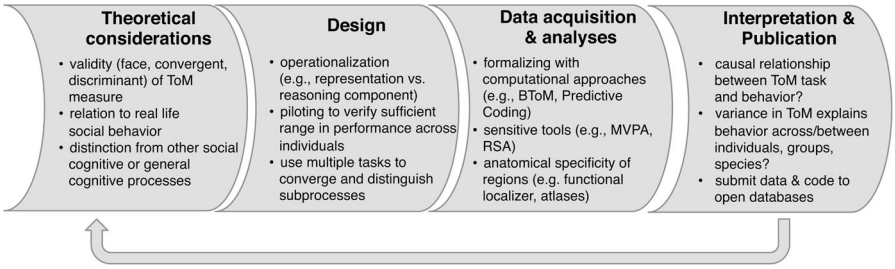

Figure 1. Summary of recommendations

Theoretical considerations should ensure construct validity and relevance to real-life social cognition. The *Design* should carefully operationalize mentalizing (ideally across multiple tasks or be precise in what narrower aspect is being investigated). During *Data Acquisition & analyses,* computational approaches can be beneficial to formalize psychological processes. Advances in analyses tools (e.g. multivariate methods) are best combined with analyses of high functional and anatomical sensitivity. The *Interpretation* should include a critical reflection on causal relationships between all specific and domain-general abilities, and on the generalizability of the results. Submitting data and specific methods to public databases (e.g., OSF, OpenfMRI) is important for reproducibility. Finally, the outcome of one scientific project should in turn fuel the next (building a cumulative science).